AD-A043 521   SYRACUSE UNIV N Y DEPT OF INDUSTRIAL ENGINEERING AND--ETC F/G 12/1
OPTIMAL CONTROL OF RANDOM WALKS.(U)
1977   R F SERFOZO   AF-AFOSR-2627-74

UNCLASSIFIED   AFOSR-TR-77-1011   NL

| OF |
AD
A043521

END
DATE
FILMED
9-77
DDC

(5)

DEPARTMENT OF INDUSTRIAL ENGINEERING AND OPERATIONS RESEARCH

See 1473

COLLEGE OF ENGINEERING
SYRACUSE UNIVERSITY
SYRACUSE, NEW YORK 13210

5

OPTIMAL CONTROL OF RANDOM WALKS

by

Richard F. Serfozo

Abstract

This is a study of a random walk on the nonnegative integers whose steps are controlled as follows. Upon arriving at a location i, a pair of probabilities (p,q) is selected from a prescribed set, a reward r(i,p,q) is received, and the next step takes the walk to locations i+1, i-1 or i, with respective probabilities p, q and 1-p-q (when i=0 these probabilities are p, 0 and 1-p). This is repeated indefinitely. A rule for successively selecting the probabilities (p,q) is a control policy. We identify conditions on the rewards and probabilities under which there exist monotonic optimal policies for discounted and average rewards. For example, in one case it is optimal to increase the probability of backward steps as the location i increases. Our results are based on (1) a criterion for monotone optimal policies, (2) a result describing when an upper envelope of concave functions is concave, and (3) a relation between optimal policies for the discounted and average reward criteria. Procedures for computing optimal policies are also presented.

Optimal Control of Random Walks

by

Richard F. Serfozo, Syracuse University

1. Introduction

We shall study a controlled random walk on the nonnegative integers that moves as follows. Upon arriving at a location i the following events occur:

(1) A pair of probabilities (p_a,q_a) is selected from the set $\{(p_1,q_1), \ldots, (p_m,q_m)\}$. Think of the (p_a,q_a), or the $a \in \{1,2, \ldots, m\}$, as the action taken. We assume that $0 \leqq p_a + q_a \leqq 1$, and at least one of these is nonzero.

(2) A real-valued reward $r(i,a)$ is received.

(3) The next location of the walk is determined by the transition probabilities

$$p(i,a,i+1) = p_a,\ p(i,a,i-1) = q_a,\ p(i,a,i\ ) = 1-p_a-q_a$$

when $i \geqq 1$; and

$$p(0,a,1) = p_a \text{ and } p(0,a,0) = 1-p_a \quad \text{when } i=0.$$

That is, the step is of size +1, -1 or 0 with respective probabilities p_a,q_a and $1-p_a-q_a$ (except at location 0). The above series of events is repeated indefinitely.

A policy f for controlling this random walk (i.e. a rule for selecting the (p_a,q_a)) is defined to be a mapping from the nonnegative integers (the state space) to $\{1,2, \ldots, m\}$ (the action space). Under the policy

This research was partially supported by Air Force Office of Scientific Research Grant AFOSR-74-2627, and NSF Grant ENG 75-13653.

f the action f(i) is taken, i.e. $(p_{f(i)},q_{f(i)})$ is selected, whenever the walk is in location i. We shall consider only these so-called stationary deterministic policies. Nothing would be gained by considering nonstatonary or randomized policies.

Each policy f, along with a rule for starting the process, determines a stochastic process $\{(X_n,a_n): n \geq 0\}$, where X_n is the location of the walk at time n, and $a_n = f(X_n)$ is the action taken. The expected discounted reward over an infinite horizon is

$$V_f(i) = E_f(\sum_{n=0}^{\infty} \alpha^n r(X_n,a_n)|X_o=i),$$

where $0 < \alpha < 1$ is a discount factor. The average reward over an infinite horizon is

$$\phi_f(i) = \underline{\lim}_{n\to\infty} n^{-1} E_f(\sum_{k=0}^{n-1} r(X_k,a_k)|X_o=i).$$

A policy f* is called α-discounted optimal if

$$V_{f*}(i) = \sup_f V_f(i) \qquad \text{for all i,}$$

and f* is called average optimal if

$$\phi_{f*}(i) = \sup_f \phi_f(i) \qquad \text{for all i.}$$

The aim is to find such policies. We shall call this decision process a controlled random walk. It is a special case of a Markov decision process, or a controlled Markov chain.

Decision processes that arise in practice often have inherent properties that lead to nicely structured optimal policies. For example, an optimal policy f(i) may be a monotone, unimodal, or convex function of i. Knowing that there is, say, an increasing optimal policy, then the search for an optimal policy may be confined to the class of increasing policies. An optimal policy might then be obtained by a simple ad hoc procedure, such as a calculus argument. This is especially important for decision processes with infinite state spaces (like ours) where optimal policies cannot be obtained by the standard procedures for processes with finite state spaces. Structured policies are also generally easier to implement than unstructured ones.

In this paper we show, under some very general conditions on the rewards r(i,a) and the probabilities (p_a, q_a), that it is (discounted and average) optimal to "increase" the probability of backward movement of the process as the location of the walk increases. We present a similar result where it is optimal to "decrease" this probability. We show how these results carry over to finite time horizons, and to walks where the set of possible probabilities for a step depends on the location where the step is taken. We also present procedures for calculating some average optimal policies.

Our analysis herein is based on three key results that apply to more general Markov decision processes. The first result is a criterion for the existence of a monotone optimal policy (Proposition 4.1). Related criteria are discussed in [6] and [8]. The second result describes when the upper envelope of a family of functions, defined on the integers, is concave (Proposition 4.2). This result enabled us to find natural conditions on the rewards r(i,a) that lead to monotone optimal policies. The third result asserts, under some weak conditions, that if a Markov decision process has a discounted optimal policy with a given structure, then it also has an average optimal policy with the same structure (Theorem 5.1). Part of this result is an extension of [2, Theorem 1].

Applications of controlled random walks arise in contexts where the descriptive theory of random walks is used. In a related paper [7], we apply the results herein to obtain optimal policies for controlling birth and death processes and queues.

2. Monotone Optimal Policies for Random Walks Based on Discounted Rewards

In this section we identify conditions under which there exist increasing and decreasing discounted optimal policies for the controlled random walk. (We use the terms increasing and decreasing to mean nondecreasing and nonincreasing, respectively.) We also discuss the monotonicity of these discounted optimal policies, with respect to the discount factor.

Here, and throughout this paper, we shall use the notation introduced above. We shall use a prime to denote the difference operator with respect to i, namely $u'(i) = u(i+1)-u(i)$. In particular, we write

$$r'(i,a) = r(i+1,a) - r(i,a).$$

Our first result concerns increasing policies. A typical increasing policy f can be written as

$$f(i) = a \qquad \text{if } i_a \leqq i < i_{a+1},$$

where $0 = i_1 \leqq i_2 \leqq \cdots \leqq i_m \leqq i_{m+1} = \infty$. This means that if the walk is in location i, and $i_a \leqq i < i_{a+1}$, then action a is taken, i.e. (p_a,q_a) is selected. Note that the action increases as i increases. Also, if $i_a = i_{a+1}$ for a particular action a, then this action is never taken.

Theorem 2.1. Suppose the following conditions hold.

(1) $p_1 \geqq p_2 \geqq \cdots \geqq p_m$, $q_1 \leqq q_2 \leqq \cdots \leqq q_m$, and $p_1 + q_m \leqq 1$.

(2) $r'(i,1) \leqq r'(i,2) \leqq \cdots \leqq r'(i,m) \leqq 0 \qquad$ for all i.

(3) $r'(i,1) \geqq r'(i+1,m) \qquad$ for all i.

Then there is an increasing α-discounted optimal policy for the random walk.

We shall prove this after we make a few observations. Theorem 2.1 asserts that there is an α-discounted optimal policy which selects higher actions in $\{1, \ldots, m\}$ as the location i of the walk increases. Under this policy, because of assumption (1), the selected q is an increasing function of i, and the selected p is a decreasing function of i. Their ratio p/q is also decreasing in i, since $p_1/q_1 \geqq \cdots \geqq p_m/q_m$. This means that the tendency of backward movement of the walk increases as its location increases. The ratio p/q is like the traffic intensity of a queueing process. We tried to prove Theorem 2.1 with (1) replaced by the weaker condition $p_1/q_1 \geqq \cdots \geqq p_m/q_m$, but we were unsuccessful. We feel that (1) cannot be relaxed this way, but we do not have a counterexample to justify this conjecture.

Note that assumption (1) poses no restriction on the (p,q)'s in the following important examples.

A Random Walk with a Controlled Ascent.

The p_a's are subscripted so that $p_1 \geqq \cdots \geqq p_m$ and $q_1 = \cdots = q_m$.

A Random Walk with a Controlled Descent.

The q_a's are subscripted so that $q_1 \leqq \cdots \leqq q_m$ and $p_1 = \cdots = p_m$.

These examples are analogous to an M/M/1 queue with a controlled arrival rate and a controlled service rate, respectively. In [7] we show that these controlled queues are actually equivalent to the above random walks, and we apply the results herein to obtain optimal policies for them.

The assumptions (1) - (3) insure that the value function of the walk (see (5)) is concave. This is a key ingredient for an increasing policy (see the verification of (8), (9) and (13), and Proposition 4.2). Note that (2) and (3) hold if and only if

$$0 \geq r'(0,m) \geq r'(0,m-1) \geq \ldots \geq r'(0,1) \geq r'(1,m) \geq r'(1,m-1) \geq \ldots \geq r'(1,1) \geq r'(2,m) \geq \ldots$$

This is a very weak restriction on the rewards. It is satisfied, for example, when

$$r(i,a) = g(a) - h(i),$$

where h is convex increasing and g has any structure. Another consequence of (2) is that the rewards are bounded from above. Namely,

$$(4) \quad \sup_{i,a} r(i,a) \leq \max_a r(0,a) < \infty.$$

We shall use the following notation and results in the proof of Theorem 2.1. We let

$$(5) \quad V(i) = \sup_f V_f(i) = \sup_f E_f\Big(\sum_{k=0}^{\infty} \alpha^k r(X_k,a_k) \Big| X_o=i\Big),$$

$$V_n(i) = \sup_f E_f\Big(\sum_{k=0}^{n-1} \alpha^k r(X_k,a_k) \Big| X_o=i\Big) \qquad \text{for } n \geq 1,$$

and $V_o(i) \equiv 0$. These are the infinite and finite horizon value functions of the random walk. Since the rewards r(i,a) are bounded from above, it follows that the V_n are finite-valued and

$$-\infty \leq V_f(i) \leq V(i) < \infty \qquad \text{for all i and f.}$$

From the theory of Markov decision processes (or dynamic programming) with upper bounded rewards, we know that the following statements hold. These come from the basic work of Bellman, Blackwell, Derman, Howard, Strauch and others, which are nicely unified and extended in [4] and [5].

(i) (Existence of Stationary Optimal Policies) An α-discounted optimal policy exists.

(ii) (Optimality Criterion) A policy f is α-discounted optimal if and only if

$$U(i,f(i)) = \max_a U(i,a) \qquad \text{for all i,}$$

where

$$U(i,a) = r(i,a) + \alpha \sum_j p(i,a,j)V(j).$$

(iii) (Optimality Equations) The V_n and V satisfy the optimality equations

$$V_n(i) = \max_a\{r(i,a) + \alpha\sum_j p(i,a,j)V_{n-1}(j)\} \qquad (n \geq 1), \text{ and}$$

$$V(i) = \max_a\{r(i,a) + \alpha\sum_j p(i,a,j)V(j)\} \qquad \text{for all } i.$$

(iv) (Value Iteration) For all i, $V(i) = \lim_{n\to\infty} V_n(i)$.

Our last preliminary for the proof of Theorem 2.1 is the following.

Lemma 2.2. If (1) - (3) hold, then $V_n(i)$ is concave decreasing in i for each $n \geq 0$.

Proof. We shall prove this by induction. Trivially, $V_o = 0$ is concave decreasing. Assume that V_n is concave decreasing. The Optimality Equations (iii) can be written

$$V_{n+1}(i) = \max_a U_n(i,a)$$

where

(6) $U_n(i,a) = r(i,a) + \alpha \sum_j p(i,a,j)V_n(j)$.

To prove that V_{n+1} is decreasing, it suffices, since V_{n+1} is the upper envelope of the functions $U_n(\cdot,1)$, ..., $U_n(\cdot,m)$, to show

(7) $U_n'(i,a) \leq 0$ for all a and i.

And to prove that V_{n+1} is concave, it suffices, by Proposition 4.2 (in Section 4), to show

(8) $U_n'(i,1) \leq U_n'(i,2) \leq \cdots \leq U_n'(i,m)$ for all i, and

(9) $U_n'(i,1) \geq U_n'(i+1,m)$ for all i.

Writing (6) in terms of the p_a and q_a we get

$$U_n(i,a) = \begin{cases} r(0,a) + \alpha[(1-p_a)V_n(0) + p_aV_n(1)] & \text{for } i = 0 \\ r(i,a) + \alpha[q_aV_n(i-1) + (1-p_a-q_a)V_n(i) + p_aV_n(i+1)] & \end{cases}$$

for $i \geq 1$.

Then for any $i \geqq 1$,

$$(10) \quad U_n'(i,a) = r'(i,a) + \alpha[q_aV_n'(i-1) + (1-p_a-q_a)V_n'(i) + p_aV_n'(i+1)]$$
$$= r'(i,a) + \alpha[V_n'(i) - q_aV_n''(i-1) + p_aV_n''(i)].$$

Under our induction hypothesis, the $V_n'(i)$ and $V_n''(i) = V_n'(i+1) - V_n'(i)$ are nonnegative. Then from the first and second lines in (10), and the assumptions (1) and (2), it follows that (7) and (8) are satisfied for $i \geqq 1$. The inequality (9) is also satisfied for $i \geqq 1$, since by (1) and (3),

$$(11) \quad U_n'(i+1,m) - U_n'(i,1) = r'(i+1,m) - r'(i,1)$$
$$+ \alpha[q_1V_n''(i-1) + (1-p_1-q_m)V_n''(i) + p_mV_n''(i+1)] \leqq 0.$$

By similar arguments it follows that (7) - (9) are also satisfied for $i = 0$. We have thus proved that V_{n+1} is concave decreasing, and this completes our induction argument.

We are now ready to prove Theorem 2.1 which asserts that (1) - (3) imply the existence of an increasing α-discounted optimal policy.

Proof of Theorem 2.1. Consider the policy

$$(12) \quad f(i) = \max\{a: U(i,a) = \max_{\tilde{a}} U(i,\tilde{a})\},$$

where

$$U(i,a) = r(i,a) + \alpha \sum_j p(i,a,j)V(j).$$

By the Optimality Criterion (ii), this f is α-discounted optimal. To complete the proof, we need only show that f is increasing. To do this it suffices, by Proposition 4.1, to show

$$(13) \quad U'(i,1) \leqq U'(i,2) \leqq \cdots \leqq U'(i,m) \qquad \text{for all } i.$$

To this end, note that

$$U(i,a) = \begin{cases} r(0,a) + \alpha[(1-p_a)V(0) + p_aV(1)] & \text{for } i = 0 \\ r(i,a) + \alpha[q_aV(i-1) + (1-p_a-q_a)V(i) + p_aV(i+1)] & \text{for } i \geqq 1. \end{cases}$$

Then

$$(14)\quad U'(i,a) = \begin{cases} r'(0,a) + \alpha[V'(0) - q_a V'(0) + p_a V''(0)] & \text{for } i = 0 \\ r'(i,a) + \alpha[V'(i) - q_a V''(i-1) + p_a V''(i)] & \text{for } i \geq 1. \end{cases}$$

By Lemma 2.2 and the Value Iteration Property (iv), it follows that V is concave. Then using (1), (2), $V'(0) \leq 0$, and $V''(i) \leq 0$ in (14), we obtain (13). This completes the proof.

We have just shown when it is optimal to increase the probability of backward movement of the random walk as its location increases. This tends to keep the walk near zero. Our next result describes the opposite situation in which it is optimal to decrease the probability of backward movement as the location increases. This tends to push the walk toward $+\infty$, accelerating its forward movement as it approaches $+\infty$. Similar results appear in [6].

Theorem 2.3. Suppose the following hold.

(15) $p_1 \geq p_2 \geq \cdots \geq p_m$ and $q_1 \leq q_2 \leq \cdots \leq q_m$.

(16) $r'(i,1) \geq r'(i,2) \geq \ldots \geq r'(i,m) \geq 0$ for all i.

(17) $r(i,a)$ is convex increasing in i for each a.

(18) $\max_a r(i,a) \leq g(i)$, where g is a polynomial function in i.

Then there is a decreasing α-discounted optimal policy for the random walk.

Note that this result does not require, as Theorem 2.1 does, that $p_1 + q_m \leq 1$. The assumptions (16) - (18) are satisfied if $r(i,a) = g_1(a) + g_2(i)$, where $g_3(i)$ is a convex increasing polynomial in i.

Proof. A sufficient condition for the above dynamic programming statements (i) - (iv) to hold, and the V_f and V to exist, is that

$$(19)\quad \limsup_{n\to\infty} \sup_f E_f\Big(\sum_{k=n}^{\infty} \alpha^k |r(X_k,a_k)| \,\Big|\, X_0 = i\Big) = 0 \qquad \text{for all } i.$$

See [5]. If the g in (18) is of the form $g(i) = i^N$, then using the fact that $P_f(X_k \leq k|X_0 = 0) = 1$, we have

$$E_f(\sum_{k=n}^{\infty} \alpha^k |r(X_k,a_k)| |X_0 = 0) \leq \sum_{k=n}^{\infty} \alpha^k k^N < \infty.$$

Similar bounds for this expected value can be obtained for any polynomial g and any value of X_0. These bounds are sufficient for (19) to hold.

Proceeding as in the proof of Theorem 2.1, we consider the policy

$$f(i) = \max\{a: U(i,a) = \max_{\tilde{a}} U(i,\tilde{a})\}.$$

This is α-discounted optimal by the Optimality Criterion. By an induction argument, as in the proof of Lemma 2.2, it follows that each n-period value function $V_n(i)$ is convex increasing in i. Here the V_n is convex increasing, since it is the upper envelope of $U_{n-1}(\cdot,1), \ldots, U_{n-1}(\cdot,m)$, which are clearly convex increasing. Then $V(i) = \lim_{n\to\infty} V_n(i)$ is convex increasing. Finally arguing as in the proof of Theorem 2.1, it follows that f is decreasing.

Our final result in this section concerns the monotonicity of α-discounted optimal policies, with respect to the discount factor α. This is of interest by itself. It is also a key result for obtaining average optimal policies from discount optimal policies, which we do in Section 6.

We shall assume here that we are dealing with a Markov decision process with transition probabilities $p(i,a,j)$, and rewards $r(i,a)$, which are bounded from above. We let

(20) $f_\alpha(i) = \max\{a: U_\alpha(i,a) = \max_{\tilde{a}} U_\alpha(i,\tilde{a})\}$,

where

$$U_\alpha(i,a) = r(i,a) + \alpha \sum_j p(i,a,j)V(i).$$

According to the Optimality Criterion, the f_α is an α-discounted optimal policy.

Theorem 2.4. If $f_\alpha(i)$ is increasing in i for each α, and

$$r(i,1) \geq r(i,2) \geq \dots \geq r(i,m), \text{ for some } i$$

then $f_\alpha(i) \leq f_\beta(i)$ for this i and all $0 \leq \alpha \leq \beta \leq 1$.

Proof. Let $b = f_\alpha(i)$. For any $\alpha \leq \beta$ it follows by the definition of f_α and the hypothesis that

$$0 < U_\alpha(i,b) - U_\alpha(i,a) = r(i,b) - r(i,a) + \alpha \sum_j [p(i,b,j) - p(i,a,j)]V(j)$$

$$\leq \alpha \sum_j [p(i,b,j) - p(i,a,j)]V(j).$$

Using this inequality we have

$$U_\beta(i,b) - U_\beta(i,a) \geq U_\alpha(i,b) - U_\alpha(i,a) > 0 \qquad \text{for } \alpha \leq \beta.$$

From this, and the assumption that $f_\beta(i)$ is increasing, we get $f_\beta(i) \geq b = f_\alpha(i)$ for $\alpha \leq \beta$. This completes the proof.

Example 2.5. Consider the controlled random walk with rewards $r(i,a) = g(a) - h(i)$, where $h(\cdot)$ is convex and increasing, and $g(\cdot)$ is decreasing. By Theorem 2.1 there is an increasing α-discounted optimal policy f_α, as defined by (20). Then by Theorem 2.4 we have $f_\alpha(i) \leq f_\beta(i)$ for all $\alpha \leq \beta$ and i.

3. Monotone Discount Optimal Policies for Random Walks with State Dependent Transitions

We have been discussing a random walk in which each step size is determined by a pair of probabilities selected from the set $\{(p_1,q_1), \dots, (p_m,q_m)\}$, where this set is independent of the location of the walk. We now consider the case where this set of probabilities is dependent on the location of the walk. We present analogs of Theorems 2.1 and 2.3.

We shall assume (only in this section) that the random walk moves as follows. Upon arriving at location i, the following events occur:

(1) A pair of probabilities $(p(i,a), q(i,a))$ is selected from the set $\{(p(i,1), q(i,1)), \ldots, (p(i,m),q(i,m))\}$.

(2) A reward $r(i,a)$ is received.

(3) The next location of the walk is determined by the transition probabilities

$$p(i,a,i+1) = p(i,a),\ p(i,a,i-1) = q(i,a),\ p(i,a,i) = 1 - p(i,a) - q(i,a),$$

when $i \geq 1$, and

$$p(0,a,1) = p(0,a) \text{ and } p(0,a,0) = 1 - p(0,a) \quad \text{when } i = 0.$$

The above series of events are repeated indefinitely.

In the following, we let

$$d(i,a) = q(i,a) - p(i,a).$$

Theorem 3.1. Suppose the following conditions hold.

(4) $p(i,1) \geq \ldots \geq p(i,m)$, $q(i,1) \leq \ldots \leq q(i,m)$ and $p(i,1) + q(i,m) \leq 1$ for all i.

(5) $d'(i,1) \leq \ldots \leq d'(i,m) \leq 0$ and $d'(i,1) \geq d'(i+1,m)$ for all i.

(6) $r'(i,1) \leq \ldots \leq r'(i,m) \leq 0$ for all i.

(7) $r'(i,1) \geq r'(i+1,m)$ for all i.

Then there is an increasing α-discounted optimal policy for the random walk.

This is similar to Theorem 2.1, except for the additional condition (5). It can be proved just as we proved Theorem 2.1. The key steps are to observe the following analogs of (10) and (11) in Section 2:

$$U_n'(i,a) = r'(i,a) + \alpha\{(1-d'(i,a))V_n'(i) - q(i,a)V_n''(i-1) + p(i,a)V_n''(i+1)\},$$

and

$$\begin{aligned} U_n'(i+1,m) - U_n'(i,1) = r'(i+1,m) - r'(i,1) + \alpha\{&q(i,1)V_n''(i-1) \\ &+ [1 - p(i+1,1) - q(i+1,m) - d'(i+1,m)]V_n''(i) \\ &+ [d'(i,1) - d'(i+1,m)]V'(i) + p(i+2,m)V_n''(i+1)\} \leq 0 \end{aligned}$$

The analog of Theorem 2.3 is as follows.

Theorem 3.2. Suppose the following conditions hold.

(8) $p(i,1) \geq \ldots \geq p(i,m)$ and $q(i,1) \leq \ldots \leq q(i,m)$ for all i.

(9) $d'(i,1) \leq \ldots \leq d'(i,m)$ for all i.

(10) $d(i,a)$ is concave decreasing in i for each a.

(11) $r'(i,1) \geq \ldots \geq r'(i,m)$ for all i

(12) $r(i,a)$ is convex in i for each a

(13) $\max_a |r(i,a)| \leq g(i)$, where g is a polynomial in i.

Then there is a decreasing α-discounted optimal policy for the random walk.

4. Criteria for Monotone Optimal Policies and Concave Value Functions

In this section we present two key results which we used above for establishing the existence of monotone optimal policies for our random walk.

We shall consider the general optimization problem

$$v(i) = \max_a u(i,a) \quad \text{for } i = 0,1,\ldots$$

where $a \in \{1, \ldots, m\}$ and u is a real-valued function. An optimal policy for this problem is defined to be any mapping f from $\{0,1,\ldots\}$ to $\{1,2,\ldots,m\}$ which satisfies

$$u(i,f(i)) = \max_{a} u(i,a) \quad \text{for all } i.$$

Note that this is an abstraction of the Optimality Criterion in dynamic programming (recall statement (ii) in Section 2).

Our first result describes sufficient conditions for the existence of monotone policies. Variations of this, along with other applications, are discussed in [6] and [8].

Proposition 4.1. Let f be the optimal policy defined by

$$f(i) = \max\{a: u(i,a) = \max_{\tilde{a}} u(i,\tilde{a})\}.$$

The optimal policy f is increasing if

(1) $u'(i,1) \le u'(i,2) \le \dots \le u'(i,m)$ for all i.

The optimal policy f is decreasing if

(2) $u'(i,1) \ge u'(i,2) \ge \dots \ge u'(i,m)$ for all i.

Proof. Suppose (1) holds, and there is an i such that $f(i+1) < f(i)$. By the definition of f(i) and (1), we have

$$0 \le u(i,f(i)) - u(i,f(i+1)) \le u(i+1,f(i)) - u(i+1,f(i+1)),$$

and so $u(i+1,f(i+1)) \le u(i+1,f(i))$. But this contradicts the definition of f(i+1). Thus f must be increasing. The assertion that (2) implies that f is decreasing is proved similarly.

In order to apply Proposition 4.1 when u(i,a) is a function of v (as we did in Section 2), some knowledge of the structure of the value function v may be required. Since v is the upper envelope of $u(\cdot,1)$, ..., $u(\cdot,m)$, then v is obviously convex, increasing or decreasing when all of the $u(\cdot,a)$'s are convex, increasing or decreasing, respectively. The next result describes conditions under which v is concave.

Proposition 4.2. The function v is concave if either of the following conditions hold.

(3) $u'(i,1) \leq u'(i,2) \leq \ldots \leq u'(i,m)$ and $u'(i,1) \geq u'(i+1,m)$ for all i.

(4) $u'(i,1) \geq u'(i,2) \geq \ldots \geq u'(i,m)$ and $u'(i,m) \geq u'(i+1,1)$ for all i.

Proof. Suppose (3) holds and let f be the optimal policy in Proposition 4.1. Using (3) and the increasing property of f we have

$$v'(i) = u(i+1,f(i+1)) - u(i,f(i)) \geq u(i+1,f(i)) - u(i,f(i))$$
$$\geq u'(i,1) \geq u'(i+1,m) \geq u(i+2,f(i+2)) - u(i+1,f(i+2)) \geq v'(i+1).$$

Thus v is concave. A similar argument shows that v is concave if (4) holds.

5. Discounted and Average Reward Optimal Policies of Similar Structure

If a Markov decision process has a discounted optimal policy with a special structure, then it seems reasonable that there should be an average optimal policy with the same structure. We shall show that this is true in a fairly general setting. In the next section we apply this to our random walk.

We shall consider a Markov decision process with rewards $r(i,a)$, and transition probabilities $p(i,a,j)$ for $i,j \geq 0$ and a in some set. We let Π denote the set of all policies f under which the α-discounted reward function $V_f(i)$ is finite-valued for all $0 < \alpha < 1$, and the limit

$$\phi_f(i) = \lim_{n\to\infty} n^{-1} E_f(\sum_{k=0}^{n-1} r(X_k,a_k)|X_0 = i)$$

exists for all i, where $-\infty \leq \phi_f(i) < \infty$.

Theorem 5.1. Suppose the Markov decision process described above has upper bounded rewards, and there is a set of policies $\Gamma = \{f_1,f_2, \ldots\}$ in Π such that f_n is α_n-discounted optimal, where α_n is a sequence with $\alpha_n \to 1$. Then

(1) $\sup_{f\in\Pi} \phi_f(i) = \sup_{f\in\Gamma} \phi_f(i)$ for all i.

If, in addition, Γ is a finite set, then there is a policy $f^* \in \Gamma$ such that

(2) $\phi_{f^*}(i) = \sup\limits_{f\in\Pi} \phi_f(i)$ for all i.

The second part of this result is a slight extension of [2, Theorem 1]. The first part is new. The usefulness of Theorem 5.1 is illustrated in the next result which follows immediately.

Corollary 5.2. If the Markov decision process in Theorem 5.1 has an increasing α-discounted optimal policy for each α, the set of such policies is finite, and $V_f(i) \equiv -\infty$ for all policies $f \notin \Pi$, then there exists an increasing average optimal policy.

Proof of Theorem 5.1. Suppose for now that the rewards $r(i,a)$ are all nonpositive. We first note that for any $f \in \Pi$,

(3) $\phi_f(i) = \lim\limits_{\alpha\to 1} (1-\alpha)V_f(i)$ for all i.

This follows by the well-known Abelian Theorem [3,p.445], when $\phi_f(i)$ is finite. And it follows when $\phi_f(i) = -\infty$, since

$$(4) \quad (1-\alpha)V_f(i) \leq (1-\alpha)E_f\Big(\sum_{k=0}^{\nu_\alpha} \alpha^k r(X_k,a_k)\,\Big|\,X_o = i\Big)$$

$$\leq \nu_\alpha^{-1}E_f\Big(\sum_{k=0}^{\nu_\alpha} \alpha^k r(X_k,a_k)\,\Big|\,X_o = i\Big) \to \phi_f(i) = -\infty \text{ as } \alpha \to 1,$$

where ν_α is the integer part of $(1-\alpha)^{-1}$.

Using (3) and the assumption that f_n is α_n-discounted optimal, we have

$$\sup_{f\in\Pi} \phi_f(i) = \sup_{f\in\Pi} \lim_{n\to\infty} (1-\alpha_n)V_f(i) \leq \lim_{n\to\infty} (1-\alpha_n)V_{f_n}(i)$$

$$\leq \sup_{f\in\Gamma} \lim_{n\to\infty} (1-\alpha_n)V_f(i) = \sup_{f\in\Gamma} \phi_f(i).$$

Furthermore, the first term in the above is always greater than or equal to this last term, and so they are equal. This proves (1).

Now assume that Γ is finite. Then there is an $f^* \in \Gamma$ which is α_{n_k}-discounted optimal for $k = 1, 2, \ldots$ where α_{n_k} is some subsequence of α_n. Using Theorem 1 in [9,p.181] (for our nonpositive rewards!) and (3), it follows for any $f \in \Pi$ that

$$\phi_f(i) \leq \underline{\lim}_{\alpha\to 1} (1-\alpha)V_f(i) \leq \underline{\lim}_{k\to\infty} (1-\alpha_{n_k})V_{f^*}(i) = \phi_{f^*}(i).$$

This proves (2).

We now prove (1) and (2) for upper bounded rewards. Let c be an upper bound for the r(i,a)'s, and consider the Markov decision process with rewards $\tilde{r}(i,a) = r(i,a) - c$, transition probabilities $p(i,a,j)$, and average rewards $\tilde{\phi}_f$. This process has the same set of α-discounted optimal policies as the original process, its rewards are nonpositive, and $\tilde{\phi}_f = \phi_f - c$. Thus, by the above

$$\sup_{f\in\Pi} \phi_f(i) = \sup_{f\in\Pi}(\tilde{\phi}_f(i) + c) = \sup_{f\in\Gamma}(\tilde{\phi}_f(i) + c) = \sup_{f\in\Gamma} \phi_f(i).$$

Now suppose Γ is finite and $f^* \in \Gamma$ is as defined in the preceeding paragraph. Then

$$\phi_{f^*}(i) = \tilde{\phi}_{f^*}(i) + c = \sup_{f\in\Pi} \tilde{\phi}_f(i) + c = \sup_{f\in\Pi} \phi_f(i) \quad \text{for all } i.$$

This completes the proof.

6. Monotone Optimal Policies for Random Walks Based on Average Rewards

Theorem 2.1 describes conditions under which there exists an increasing α-discounted optimal policy for the random walk. In this section, we show that these condition, with some minor additions, are also sufficient for the existence of an increasing average optimal policy. A similar result holds for decreasing average optimal policies (based on Theorem 2.3), but for the sake of brevity, we shall not discuss it.

We shall consider the random walk as described in Sections 1 and 2. In our first result, we use the following conditions.

(1) $p_1 \geq \cdots \geq p_m$, $q_1 \leq \cdots \leq q_m$, and $p_1 + q_m \leq 1$.

(2) $p_1 > 0$, $q_1 > 0$ and $p_m/q_m < 1$.

(3) $r'(i,1) \leq \cdots \leq r'(i,m) \leq 0$ for all i, and at least one of the $r'(i,a)$ is nonzero.

(4) $r'(i,1) \geq r'(i+1,m)$ for all i.

Under these assumptions the average reward

$$\phi_f(i) = \lim_{n\to\infty} n^{-1} E_f\Big(\sum_{k=0}^{n-1} r(X_k,a_k) \Big| X_o = i\Big)$$

exists for any policy f and $-\infty \leq \phi_f(i) < \infty$ (see Proposition 6.2). Moreover, $\phi_f(i)$ is independent of i, so we shall simply denote it by ϕ_f. We let I denote the set of increasing α-discounted optimal policies for $0 < \alpha < 1$. Such policies exist under (1) - (4), by Theorem 2.1.

Theorem 6.1. Suppose the random walk satisfies (1) - (4). Then

$$\sup_f \phi_f = \sup_{f \varepsilon I} \phi_f.$$

If, in addition, there is an N such that

$$r(i,1) \geq r(i,2) \geq \cdots \geq r(i,m) \qquad \text{for all } i \geq N,$$

then there is an increasing policy in I that is average optimal for the random walk.

The first assertion says that the increasing policies in I yield the largest average reward, but it doesn't say that one of the policies in I actually attains the maximum reward. The second assertion does. The assumptions (1) - (4) are essentially assumptions (1) - (3) in

Theorem 2.1 with a few minor additions. These additions simply eliminate some degenerate cases. Specifically, we assume that at least one of the $r'(i,a)$ is nonzero to rule out the case where the rewards do not depend on i. With this case ruled out, (3) and (4) imply that $r(i,a) \downarrow -\infty$ as $i \to \infty$ for all a. We assume $q_1 > 0$ for the sake of brevity. The analysis presented here also carries over to the case when some of the q_a's are zero, but more details are involved. The $p_1 > 0$, in conjunction with $q_1 > 0$, just eliminates the case in which $p_1 = \ldots = p_m = 0$, and each policy determines a walk that is absorbed at zero. Even though $p_1 > 0$, some of the other p_a's may be zero. The $p_m/q_m < 1$, eliminates the case in which each policy f determines a walk whose states are all transient or null recurrent, and whose average reward $\phi_f = -\infty$ (see Proposition 6.2). Here any policy is average optimal.

Proof. If $\phi_f = -\infty$ for all f, then the assertions are trivally satisfied.

Now suppose there is an f_o with $\phi_{f_o} \neq -\infty$. It follows that its α-discounted reward $V_{f_o}(i) > -\infty$ for all i and α. Consequently, for each $f \in I$ we have $V_f(i) \geq V_{f_o}(i) > -\infty$ for all i. In addition, $\phi_f > -\infty$ for each $f \in I$. For if not, then arguing as in (4) in Section 5, we would have $V_f(i) = -\infty$. Let Π be the set of policies f for which $\phi_f > -\infty$. Then from Proposition 6.2, Theorem 2.1, and Theorem 5.1, it follows that

$$\sup_f \phi_f = \sup_{f \in \Pi} \phi_f = \sup_{f \in I} \phi_f .$$

We now establish the existence of an increasing average optimal policy. Let $\bar{a}$ denote the largest action taken by any policy in I. Let f be a policy in I that takes the action $\bar{a}$, and let α_o be a discount factor such that f is α_o-discounted optimal. Let Γ be the set of increasing α-discounted policies f_α, as constructed in the proof of

Theorem 2.1, for all $\alpha_o \leqq \alpha < 1$. Then Theorem 2.4 yields

$$f_\alpha(i) = \bar{a} \quad \text{for all } i \geqq \min\{j \geqq N\colon f(j) = \bar{a}\} \text{ and } \alpha > \alpha_o.$$

Consequently, Γ is a finite set. Thus from Theorem 5.1 it follows that there is an increasing policy in I that is average optimal for the random walk.

In the next result we present an expression for the average reward function ϕ_f of our random walk. We used this in the above proof and we shall use it again in the next section.

Proposition 6.2. Suppose the random walk satisfies (1), (2) and $r(i,a) \downarrow -\infty$ as $i \to \infty$ for all a. Let f be a policy, and let

$$\gamma_i = \prod_{k=0}^{i-1} p_{f(k)}/q_{f(k+1)} \qquad \text{for } i \geqq 1,$$

and

$$N = \begin{cases} \infty & \text{if } p_{f(i)} > 0 \text{ for all } i \\ \min\{i \geqq 0\colon p_{f(i)} = 0\} & \text{otherwise.} \end{cases}$$

Case 1. The random walk under f is such that $\{0,1,\ldots,N\}$ is a closed class of positive recurrent states and $\{N+1,N+2,\ldots\}$ are transient states if and only if $\sum_i \gamma_i < \infty$. (When $N = \infty$ the latter set is null.) In this case, the limiting distribution of the walk is

$$\pi_f(i) = \begin{cases} \gamma_i & \text{if } i \geqq 1 \\ (1 + \sum_{i=1}^{\infty} \gamma_i)^{-1} & \text{if } i = 0, \end{cases}$$

(if $N < \infty$, then $\pi_f(i) = \gamma_i = 0$ for $i > N$), and

$$\phi_f(i) = r(0,f(0))\pi_f(0) + \sum_{k=1}^{N} r(k,f(k))\gamma_k \qquad \text{for all } i.$$

Case 2. The random walk under f has all transient or null recurrent states if and only if $\sum_{i=1}^{\infty} \gamma_i = \infty$. In this case $\phi_f(i) = -\infty$ for all i.

Proof. The assertions concerning the classification of states follow by standard arguments for Markov chains. For the case in which $\sum_i \gamma_i < \infty$, we have

$$\phi_f(i) = \sum_{k=0}^{\infty} r(k,f(k))\pi_f(k) = r(0,f(0))\pi_f(0) + \sum_{k=1}^{N} r(k,f(k))\gamma_k.$$

See [1,Corollary 6.2.23.] In this reference the assumption that r(k,f(k)) is bounded can be relaxed: one only needs that the above sum exists (± ∞ being possible values). In our context, $-\infty \leqq \phi_f(i) < \infty$.

It remains to prove the last assertion. To this end, assume that $\sum_i \gamma_i = \infty$. Let $\nu_i(n)$ denote the number of visits that the random walk, under f, makes to state i in n steps. The r(i,a) is decreasing in i, and so for each j > 1,

$$n^{-1} \sum_{k=0}^{n} r(X_k,a_k) = n^{-1} \sum_{i=0}^{j-1} \nu_i(n) r(i,f(i)) + n^{-1} \sum_{i=j}^{\infty} \nu_i(n) r(i,f(i))$$

$$\leqq r(0,f(0))n^{-1} \sum_{i=0}^{j-1} \nu_i(n) + r(j,f(j))n^{-1} \sum_{i=j}^{\infty} \nu_i(n)$$

$$= r(j,f(j)) + n^{-1} \sum_{i=0}^{j-1} \nu_i(n)[r(0,f(0)) - 1].$$

Since each i is transient or null recurrent, then $n^{-1}\nu_i(n) \to 0$ a.s. It follows that

$$\overline{\lim_{n\to\infty}}\, n^{-1} E_f\left(\sum_{k=0}^{n-1} r(X_k,a_k) \,|\, X_o = i\right) \leqq r(j,f(j))$$

Letting $j \to \infty$ yields $\phi_f(i) = -\infty$ for all i.

7. Computation of Optimal Average Reward Policies: The Two Action Case

We shall show how to compute an increasing average optimal policy for a random walk with two control action (p_1,q_1) and (p_2,q_2). We discuss the multi-action case in the next section.

In addition to the notation in Sections 1,2 and 6, we let $\rho_a = p_a/q_a$ and let

$$D_n = -(\rho_1 - \rho_2)(1 - \rho_2)^{-1}\sum_{i=0}^{n-1} r(i,1)\rho_1^{\,i} + \rho_1[r(n,1) - r(n,2)]$$

$$[\rho_1^{\,n-1}\rho_2(1-\rho_2)^{-1} + \sum_{i=0}^{n-1} \rho_1^{\,i}] + (\rho_1 - \rho_2)(\sum_{i=0}^{n-1} \rho_1^{\,i})\sum_{i=n}^{\infty} r(i,2)\rho_2^{\,i-n}.$$

For each $n (0 \leqq n \leqq \infty)$, we define the policy

$$f_n(i) = \begin{cases} 1 & \text{if } 0 \leqq i < n \\ 2 & \text{if } i \geqq n. \end{cases}$$

Note that $f_\infty(i) \equiv 1$ and $f_o(i) \equiv 2$. We also let

$$n^* = \begin{cases} \infty & \text{if } D_n > 0 \quad \text{for all } n \geqq 0 \\ \min\{n \geqq 0\colon D_n \leqq 0\} & \text{otherwise.} \end{cases}$$

Theorem 7.1. Suppose the random walk with two actions satisfies the following conditions.

(1) $0 < q_1 \leqq q_2$, $\rho_1 > \rho_2$, $\rho_2 < 1$, and $p_1 + q_2 \leqq 1$.

(2) $r'(i,1) \leqq r'(i,2) \leqq 0$ for all i, and $r'(i,a) \neq 0$ for some a and i.

(3) $r'(i,1) \geqq r'(i+1,2)$ for all i.

(4) $\sum_{i=0}^{\infty} r(i,2)\rho_2^{\,i} > -\infty$.

Then the policy f_{n^*} is average optimal.

This says that it is average optimal to select (p_1,q_1) when the walk is below n^* and to select (p_2,q_2) otherwise. The n^* can be obtained when $r(i,2)$ is tractable, say a polynomial in i, by successively computing $D_1, D_2, \ldots$ until a D_n is reached such that $D_n \leqq 0$. Then $n^* = n$. The case when $r(i,a) = g(a) - hi$ is discussed below. In the proof of Theorem 7.1 we show that D_n is decreasing in n. This can be used in an obvious way to shorten the procedure for obtaining the n*.

Theorem 7.1 is essentially a special case of Theorem 6.1. We added assumption (4) here, because without it, any increasing policy is average optimal. This follows from Theorem 6.1 and the fact that the average reward for each increasing policy is $-\infty$, as seen from (8) and (9) below.

Proof of Theorem 7.1. Let ϕ_n denote the average reward of the walk under the policy $f_n (0 \leq n \leq \infty)$. From Theorem 6.1 it follows that

$$\sup_f \phi_f = \sup_n \phi_n .$$

Then in order to prove that the policy f_{n*} is average optimal it suffices to show

$$(6) \quad \phi_{n*} = \sup_n \phi_n .$$

To this end, we first note that by Proposition 6.2, the limiting distribution $\pi_n(\cdot)$ of the walk under policy f_n is as follows:

$$\pi_\infty(i) = (1-\rho_1)\rho_1^{\,i} \qquad \text{for } i \geq 0,$$

and for $n \geq 0$,

$$\pi_n(i) = \begin{cases} \pi_n(0)\rho_1^{\,i} & \text{for } 0 \leq i < n \\ \pi_n(0)\rho_1^{\,n-1}\rho_2^{\,i-n+1} & \text{for } i \geq n, \end{cases}$$

where

$$\pi_n(0) = \Big[\sum_{k=0}^{n-1} \rho_1^{\,k} + \rho_1^{\,n-1}\rho_2(1-\rho_2)^{-1} \Big]^{-1} .$$

(Here $\sum_{k=0}^{-1} = 0$.) Another application of Proposition 6.2, using the above π_n's, yields

(7) $\phi_\infty = (1-\rho_1) \sum_{i=0}^{\infty} r(i,1)\rho_1^{\,i}$

and for $n \geq 0$,

(8) $\phi_n = \pi_n(0) \left\{ \sum_{i=0}^{n-1} r(i,1)\rho_1^{\,i} + \rho_1^{\,n-1} \sum_{i=n}^{\infty} r(i,2)\rho_2^{\,i-n+1} \right\}.$

Note that

(9) $\phi_\infty = \lim_{n\to\infty} \phi_n.$

We now show that ϕ_n has a global maximum. Using (8) we have

$$(10)\quad \phi_{n+1} - \phi_n = \pi_n(0)\pi_{n+1}(0)\{(\pi_n(0)^{-1} - \pi_{n+1}(0)^{-1}) \sum_{i=0}^{n-1} r(i,1)\rho_1^{\,i}$$

$$+ \rho_1^{\,n}\pi_n(0)^{-1}(r(n,1) - r(n,2))$$

$$+ \rho_1^{\,n-1}(\rho_1\pi_n(0)^{-1} - \rho_2\pi_{n+1}(0)^{-1}) \sum_{i=n}^{\infty} r(i,2)\rho_2^{\,i-n}\}.$$

From the above expression for $\pi_n(0)$, we obtain

$$\pi_n(0)^{-1} - \pi_{n+1}(0)^{-1} = \rho_1^{\,n-1}(\rho_2 - \rho_1)/(1 - \rho_2),$$

and

$$\rho_1\pi_n(0)^{-1} - \rho_2\pi_{n+1}(0)^{-1} = (\rho_1 - \rho_2) \sum_{k=0}^{n-1} \rho_1^{\,k}.$$

Using these in (10) yields

(11) $\phi_{n+1} - \phi_n = \rho_1^{\,n-1}\pi_n(0)\pi_{n+1}(0)D_n,$

where D_n is defined above. In light of this factorization, the ϕ_n will have a global maximum if D_n is decreasing in n. From the definition of D_n and some algebraic manipulations we can write

$$(12)\quad D_{n+1} - D_n = \rho_1[\pi_n(0)^{-1} - \pi_{n+1}(0)^{-1}]r(n,1) + \rho_1\pi_{n+1}(0)^{-1}[r(n+1,1) - r(n+1,2)]$$

$$- \rho_1\pi_n(0)^{-1}[r(n,1) - r(n,2)] + (\rho_1 - \rho_2)[\sum_{k=0}^{n-1} \rho_1{}^k(1 - \rho_2) + \rho_1{}^n]$$

$$\sum_{i=n+1}^{\infty} r(i,2)\rho_2{}^{i-n-1} - [\rho_1\pi_n(0)^{-1} - \rho_2\pi_{n+1}(0)^{-1}]r(n,2)$$

$$= \pi_{n+1}(0)^{-1}\{\rho_1 r'(n,1) - \rho_2 r'(n,2)$$

$$(\rho_1-\rho_2)[(1-\rho_2)\sum_{i=n+1}^{\infty} r(i,2)\rho_2{}^{i-n-1} - r(n+1,2)]\}.$$

Under our assumption (2), we have $r'(n,1) \le r'(n,2)$. And for $i \ge n + 1$,

$$r(i,2) \le r(n+1,2) + (i-n-1)r'(n+1,2),$$

so that

$$\sum_{i=n+1}^{\infty} r(i,2)\rho_2{}^{i-n-1} \le r(n+1,2)(1 - \rho_2)^{-1} + \rho_2(1 - \rho_2)^{-2}r'(n+1,2).$$

Using these expressions in (12) yields

$$D_{n+1} - D_n \le \pi_{n+1}(0)^{-1}(\rho_1 - \rho_2)r'(n,2)(1 - \rho_2)^{-1} \le 0.$$

This says that D_n is decreasing, and so from (11) we know that ϕ_n has a global maximum.

Suppose that $D_n > 0$ for all n. Then ϕ_n is increasing, and recall that $n^* = \infty$. Thus from (9) we get

$$\phi_{n^*} = \phi_\infty = \sup_n \phi_n.$$

Now suppose $D_n \le 0$ for some n. Here the ϕ_n increases until n reaches $n^* = \min\{n\colon D_n \le 0\}$, and then it decreases to ϕ_∞, because of (9). Consequently, $\phi_{n^*} = \sup_n \phi_n$. Since these two cases cover all possibilities, if follows that the policy f_{n^*} is average optimal.

We now consider a special case of the preceding result.

Corollary 7.2. Suppose the random walk with two actions satisfies the following conditions.

(13) $0 < q_1 \leqq q_2,\ \rho_1 > \rho_2,\ \rho_2 < 1$ and $p_1 + q_2 \leqq 1$.

(14) $r(i,a) = g(a) - hi$, where $g(1) > g(2)$ and $h > 0$.

Then an average optimal policy is to select (p_1,q_1) when the walk is below n^*, and select (p_2,q_2) otherwise. The n^* is the smallest non-negative integer n for which $\tilde{D}_n \geqq 0$, where

$$\tilde{D}_n = \begin{cases} n + c\rho_1^{\,n} + c - \rho_2(g(1) - g(2))/(h(\rho_1 - \rho_2)(1 - \rho_2)) & \text{if } \rho_1 \neq 1 \\ n^2 + n(1 + \rho_2)/(1 - \rho_2) - 2\rho_2(g(1) - g(2))/(h(\rho_1 - \rho_2)) & \text{if } \rho_1 = 1, \end{cases}$$

and $c = (\rho_1 - \rho_2)/((1 - \rho_1)(1 - \rho_2))$. Furthermore,

$$\text{(16)}\quad n^* \leqq \begin{cases} \rho_2(g(1) - g(2))/(h(\rho_1 - \rho_2)(1 - \rho_1)) & \text{if } \rho_1 \neq 1 \\ [2\rho_2(g(1) - g(2))/(h(\rho_1 - \rho_2))]^{1/2} & \text{if } \rho_1 = 1. \end{cases}$$

Proof. By Theorem 7.1 it follows that an average optimal policy is to select (p_1,q_1) when the walk is below n^* and select (p_2,q_2) otherwise. Here

$$n^* = \begin{cases} \infty & \text{if } D_n > 0 \text{ for all } n \\ \min\{n\colon D_n \leqq 0\} & \text{otherwise,} \end{cases}$$

where

$$\begin{aligned} D_n = {} & -(\rho_1 - \rho_2)(1 - \rho_2)^{-1}[g(1)z_n - h\sum_{i=0}^{n-1} i\rho_1^{\,i}] \\ & + \rho_1(g(1) - g(2))[\rho_2\rho_1^{\,n-1}(1 - \rho_2)^{-1} + z_n] \\ & + z_n(\rho_1 - \rho_2)[g(2)(1 - \rho_2)^{-1} - h\sum_{i=n}^{\infty} i\rho_2^{\,i-n}], \end{aligned}$$

and $z_n = \sum_{i=0}^{n-1} \rho_1^{\,i}$.

We shall now show that $n^* = \min\{n: \tilde{D}_n \geq 0\}$. Using the identities

$$\sum_{i=n}^{\infty} i\rho_2^{\,i-n} = \rho_2 \sum_{i=n}^{\infty} (i-n)\rho_2^{\,i-n-1} + n \sum_{i=n}^{\infty} \rho_2^{\,i-n} = \rho_2(1-\rho_2)^{-2} + n(1-\rho_2)^{-1},$$

and

$$\begin{aligned} \sum_{i=0}^{n-1} i\rho_1^{\,i} &= [1 - \rho_1^{\,n} - n(1-\rho_1)\rho_1^{\,n}]/(1-\rho_1) && \text{if } \rho_1 \neq 1 \\ &= n(n-1)/2 && \text{if } \rho_1 = 1, \end{aligned}$$

in the above expression for D_n, yields

$$\begin{aligned} D_n &= \rho_2(g(1) - g(2))(1-\rho_2)^{-1} + h(\rho_1-\rho_2)(1-\rho_2)^{-1}\Big[\sum_{i=0}^{n-1} i\rho_1^{\,i} - z_n\rho_2(1-\rho_2)^{-1} + n\Big] \\ &= -h(\rho_1-\rho_2)(1-\rho_1)^{-1}(1-\rho_2)^{-1}\tilde{D}_n && \text{if } \rho_1 \neq 1 \\ & \quad\, -(1/2)h(\rho_1-\rho_2)(1-\rho_2)^{-1}\tilde{D}_n && \text{if } \rho_1 = 1. \end{aligned}$$

Note that $\tilde{D}_n$ is strictly increasing and eventually becomes positive. Consequently,

$$n^* = \min\{n: D_n \leq 0\} = \min\{n: \tilde{D}_n \geq 0\}.$$

The n^* is bounded as indicated in (16). This follows since $\tilde{D}_n$ is strictly increasing, and clearly $\tilde{D}_n > 0$ when n equal or exceeds the right size of (16).

8. A Linear Program for Computing Average Optimal Policies

The random walk we have been studying has an infinite state space. Therefore, we cannot compute optimal policies for it by the standard linear programming or policy improvement procedures for finite state processes. When the rewards $r(i,a)$ are nice functions of i (like polynomials), then the average reward ϕ_f for a monotone policy f is tractable (recall Proposition 6.2), and optimal policies might be obtainable via policy improvement. It is sometimes feasible to compute monotone optimal policies directly from the function ϕ_f, for a small number of actions. We actually did this for two actions in the last section. In this section, we discuss another approach for computing monotone average optimal policies. This is similar to the linear programming approach for finite state processes.

We shall consider a random walk (such as in Theorem 6.1) which has an increasing average optimal policy f. We assume the following:

Boundedness Assumption. There is an N such that $f(i) = m$ for all $i \geq N$ (i.e. it is average optimal to select (p_m,q_m) when the walk is in locations $i \geq N$).

We will discuss this below. We also assume, for simplicity, that $p_a > 0$ and $q_a > 1/2$ for all a. This insures that each policy determines a positive recurrent random walk.

We let $\pi = \{\pi(i,a)\colon i \geq 1,\ 1 \leq a \leq m,\ \pi(i,m) = 1 \quad \text{for } i \geq N\}$ denote a randomized policy; the $\pi(i,a)$ is the probability of selecting action a when the walk is at location i, and action m is selected for all $i \geq N$. Under the policy π, the Markov chains $\{(X_n,a_n)\}$ and $\{X_n\}$ are positive recurrent. Letting

$$\nu(i,a) = \lim_{n\to\infty} P_\pi(X_n = i,\ a_n = a \mid X_o = j),$$

the average reward is

$$(1) \quad \phi_\pi = \sum_{i=0}^{\infty} \sum_{a=1}^{m} \nu(i,a)r(i,a).$$

We can write

$$(2) \quad \nu(i,a) = \nu(i)\pi(i,a) \text{ and } \nu(i) = \sum_{a=1}^{m} \nu(i,a)$$

where

$$\nu(i) = \lim_{n\to\infty} P_\pi(X_n = i \mid X_o = j).$$

Since $\pi(i,m) = 1$ for $i \geq N$, then by Proposition 6.2 it follows that

(3) $\nu(i) = \nu(N)(p_m/q_m)^{i-N}$ for $i \geq N$.

Consequently, expression (1) simplifies to

$$\phi_\pi = \sum_{i=0}^{N-1} \sum_{a=1}^{m} \nu(i,a) r(i,a) + c(N) \sum_{a=1}^{m} \nu(N,a)$$

where

(4) $c(N) = \sum_{k=0}^{\infty} r(N+k,m)(p_m/q_m)^k$.

The problem of maximizing the ϕ_π over π is clearly equivalent to the following linear programming problem:

$$\max_{\nu(i,a)} \sum_{i=0}^{N-1} \sum_{a=1}^{m} \nu(i,a) r(i,a) + c(N) \sum_{a=1}^{m} \nu(N,a)$$

subject to

$$\sum_{a=1}^{m} \nu(0,a) = \sum_{a=1}^{m} [\nu(0,a)(1-q_a) + \nu(1,a)q_a]$$

$$\sum_{a=1}^{m} \nu(j,m) = \sum_{a=1}^{m} [\nu(j-1,a)p_a + \nu(j,a)(1-p_a-q_a) + \nu(j+1,a)q_q]$$

for $1 \leq j \leq N-1$,

$$\sum_{a=1}^{m} \nu(N,m) = \sum_{a=1}^{m} [\nu(N-1,a)p_a + \nu(N,a)(1-p_a-q_a) + \nu(N,a)(p_m/q_m)q_a]$$

$$\sum_{i=0}^{N-1} \sum_{a=1}^{m} \nu(i,a) + (1 - p_m/q_m)^{-1} \sum_{a=1}^{m} \nu(N,a) = 1$$

$$0 \leq \nu(0,a) \leq \ldots \leq \nu(N,a) \leq 1 \qquad \text{for } 1 \leq a \leq m.$$

Note that the constraints imply that $\nu(i,a)$ is the limiting distribution of $\{(X_n,a_n)\}$. An optimal solution $\nu(i,a)$ of the linear program, determines (using (2)) a monotone average optimal policy

$$\pi(i,a) = \nu(i,a)\left(\sum_{a=1}^{m} \nu(i,a)\right)^{-1} \qquad 0 \leq i \leq N-1$$

$$\pi(i,m) = 1 \qquad i \geq N.$$

This optimal policy will be a nonrandom policy when the $\nu(i,a)$ is calculated by the simplex algorithm, since a nonrandom optimal policy exists. Note that the reason we could reduce our problem to a finite variable problem is that the limiting distribution $\nu(i)$ of our random walk satisfies (2).

If the Boundedness Assumption does not hold, then the above procedure is still useful. It may not yield a truly optimal policy, but it will yield a suboptimal policy that maximizes ϕ_π over all increasing policies π which select action m for all $i \geq N$. Such a policy, when N is large, should be close to being optimal.

We initially thought that the Boundedness Assumption could be justified as follows. Consider a random walk with two actions $(\tilde{p}_1,\tilde{q}_1) = (p_{m-1},q_{m-1})$, $(\tilde{p}_2,\tilde{q}_2) = (p_m,q_m)$, and rewards $\tilde{r}(i,1) = r(i,1)$ and $\tilde{r}(i,2) = r(i,m)$, where the probabilities and rewards on the left of the equalities are from the random walk with m actions. Suppose that an average optimal policy for this two action problem is to select $(\tilde{p}_1,\tilde{q}_1)$ in location below $\tilde{n}^*$ and select $(\tilde{p}_2,\tilde{q}_2)$ otherwise. (The $\tilde{n}^*$ could be calculated as in the previous section.) Because of the way we defined the two action walk, it appears that n^* could be used as N in the Boundedness Assumption. We tried very hard to prove this, but we could not.

9. Monotone Optimal Policies for Finite Time Horizons

The above analysis for random walks over an infinite time horizon can also be done for walks over a finite time horizon. To illustrate this, we shall present a finite time horizon analogue of Theorem 2.1.

We shall consider the random walk as in Sections 1 and 2 for N time periods. Nonstationary rewards and policies are of interest for finite horizons. Accordingly, we let $r_n(i,a)$ denote the reward if the walk is in location i at time n and action a is taken. (In Section 2 this was $\alpha^n r(i,a)$.) A policy is a sequence $\underset{\sim}{f} = (f_1, \ldots, f_N)$ of mappings from the state space $\{0, 1, \ldots\}$ to the action space $\{1, \ldots, m\}$ with the interpretation that action $f_n(i)$ is taken, i.e. $(p_{f_n(i)}, q_{f_n(i)})$ is selected, if the process is in state i at time n. We let

$$V_{n,\underset{\sim}{f}}(i) = E_{\underset{\sim}{f}}\left(\sum_{k=N-n}^{N} r_k(X_k, a_k) \mid X_{N-n} = i \right)$$

and

$$V_n(i) = \sup_{\underset{\sim}{f}} V_{n,\underset{\sim}{f}}(i).$$

A policy $\underset{\sim}{f}^*$ is called optimal if

$$V_{n,\underset{\sim}{f}^*}(i) = V_n(i) \qquad \text{for all } i.$$

The Optimality Criterion for finite time horizons asserts that a policy $\underset{\sim}{f}$ is optimal if and only if

$$U_n(i, f_n(i)) = \max_a U_n(i,a),$$

where

$$U_n(i,a) = r_n(i,a) + \sum_j p(i,a,j) V_{n-1}(j)$$

and V_o is the zero function. We shall consider the optimal policy $\underset{\sim}{f}$ defined by

$$f_n(i) = \max\{a: U_n(i,a) = \max_{\tilde{a}} U_n(i,\tilde{a})\}.$$

Theorem 9.1. Suppose the following conditions hold.

(1) $p_1 \geq \cdots \geq p_m$, $q_1 \leq \cdots \leq q_m$, and $p_1 + q_m \leq 1$.

(2) $r'_n(i,1) \leq \cdots \leq r'_n(i,m)$ for all i and n.

(3) $r'_n(i,1) \geq r'_n(i+1,m)$ for all i and n.

Then $f_n(i)$ is increasing in i for each n.

Proof. This can be proved as we proved Theorem 2.1.

References

[1] Cinlar, E. (1975) Introduction to Stochastic Processes. Prentice-Hall, New Jersey.

[2] Derman, C. (1962) On sequential decisions and Markov chains. Management Science 9, 16-24.

[3] Feller, W. (1971) Introduction to Probability and Its Applications Vol. II, 2nd Ed. John Wiley, New York.

[4] Hinderer, K. (1970) Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter. Lecture Notes in Operations Research and Mathematical Sciences #33, Springer-Verlag, New York.

[5] Schal, M. (1975) Conditions for optimality in dynamic programming for the limit of n-stage optimal policies to be optimal. Z. Wahrscheinlichkeitstheorie verw. Geb. 32, 179-196.

[6] Serfozo, R. (1977) Monotone optimal policies for Markov decision processes. Stochastic Systems: Modeling Identification and Optimization, II. Mathematical Programming Study 6, 202-216, North-Holland, New York.

[7] Serfozo, R. (1977) Optimal control of birth and death processes and queues. Technical Report, IE & OR Department, Syracuse University.

[8] Topkis, D. (1975) Applications of minimizing a subadditive function on a lattice. Technical report. Hebrew University, Jerusalem.

[9] Widder, D. (1941) The Laplace Transform. Princeton University Press, Princeton, New Jersey.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

18 19 REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
AFOSR-TR-77-1011

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
6 OPTIMAL CONTROL OF RANDOM WALKS.

5. TYPE OF REPORT & PERIOD COVERED
9 Interim rept.,

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
10 Richard F./Serfozo

8. CONTRACT OR GRANT NUMBER(s)
15 AF-AFOSR 74-2627-74,
NSF-ENG-75-13653

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Syracuse University
Dept of Industrial Eng & Operations Research
Syracuse, NY 13210

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
16 17
61102F/2304/A5

11. CONTROLLING OFFICE NAME AND ADDRESS
Air Force Office of Scientific Research/NM
Bolling AFB, Washington, DC 20332

12. REPORT DATE
11 1977

13. NUMBER OF PAGES
32

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
12 38p.

15. SECURITY CLASS. (of this report)
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This is a study of a random walk on the nonnegative integers whose steps are controlled as follows. Upon arriving at a location i, a pair of probabilities (p,q) is selected from a prescribed set, a reward r(i,p,q) is received, and the next step takes the walk to locations i+1, i-1, or i, with respective probabilities p, q and 1-p-q (when i=0 these probabilities are p, 0 and 1-p). This is repeated indefinitely. A rule for successively selecting the probabilities (p,q) is a control policy~~. We identify~~ conditions are identified on the rewards and probabilities under which there exist monotonic optimal policies for → next page

20. Abstract

discounted and average rewards. For example, in one case it is optimal to increase the probability of backward steps as the location i increases. Our results are based on (1) a criterion for monotone optimal policies, (2) a result describing when an upper envelope of concave functions is concave, and (3) a relation between optimal policies for the discounted and average reward criteria. Procedures for computing optimal policies are also presented.